HW#5

Jiale Shi

1 Factor analysis

A. One problem with mixture models is that they only use a single latent variable to generate the observations. One model that addresses this issue is the Factor analysis. FA is a low rank parameterization of an MVN. In this context,

(a) Derive an expression for the number of independent parameters in the factor analysis model.

(b) Show that the factor analysis model is invariant under rotations of the latent space coordinates.

B. We consider a data set of $D = 11$ variables and $N = 387$ cases describing various aspects of cars, such as the engine size, the number of cylinders, the miles per gallon (MPG), the price, etc. Fit a $L = 2$ dimensional factor analysis model. Plot the scores in $\mathbb{R}^2$ to visualize the results. To get a better understanding of the "meaning" of the latent factors, project unit vectors corresponding to each of the feature dimensions $\boldsymbol{e}_1 = (1, 0, \ldots, 0)$, $\boldsymbol{e}_2 = (0, 1, \ldots, 0)$ etc into the low dimensional space. Represent using **biplot**. Provide your discussion.

The required data set can be downloaded from this link.

A. FA can be thought of as away of specifying a joint density model on x using a small number of parameters.

$$P(x_i|\theta) = \int N(x_i|Wz_i+\mu, \Psi) N(z_i|\mu_0, \Sigma_0)\, dz_i$$

$$= N(x_i|W\mu_0+\mu, \Psi+W\Sigma_0 W^T)$$

We can set $\mu_0 = 0$ without a loss of generality

We can absorb $W\mu_0$ into μ, $W\mu_0 + \mu \longrightarrow \mu$

We can set $\Sigma_0 = I$

We can emulate a correlated prior by defining a new weight matrix $\tilde{W} = W\Sigma_0^{\frac{1}{2}}$

$$\text{cov}[x|\theta] = W\Sigma_0 W^T + \Psi$$

$$= (\tilde{W}\Sigma_0^{-\frac{1}{2}})\Sigma_0(\tilde{W}\Sigma_0^{-\frac{1}{2}})^T + \Psi$$

$$= \tilde{W}\tilde{W}^T + \Psi$$

We thus see that FA approximates the covariance matrix of visible vector using a low-rank decomposition

$$C \triangleq \text{cov}[x] = WW^T + \Psi$$

W only uses $O(KD)$ parameters, which allows a flexible compromise between a full covariance Gaussian with $O(D^2)$ and a diagonal covariance with $O(D)$ parameters.

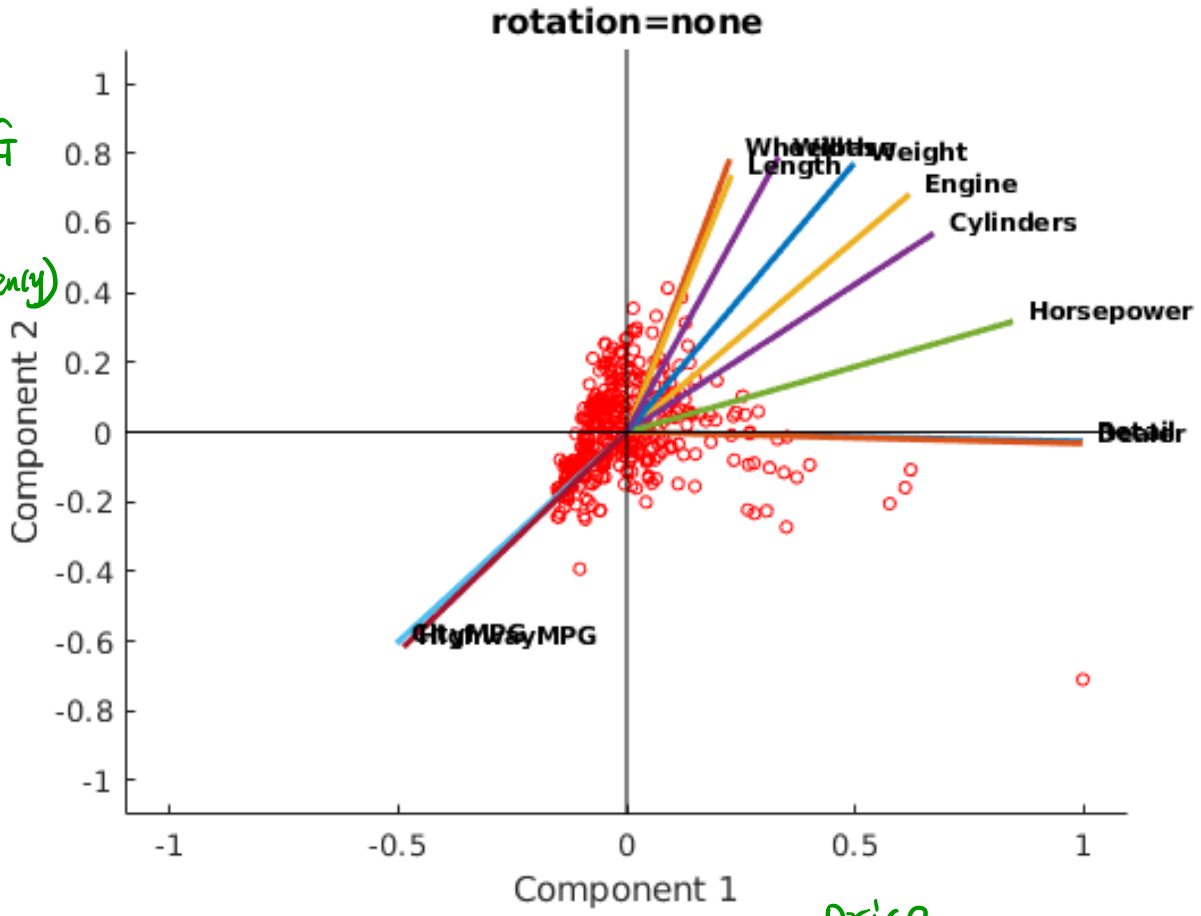
(b). Consider an arbitrary orthogonal rotation matrix R, satisfying $RR^T = I$. Let us define $\tilde{W} = WR$

$$\tilde{W}W^T = WRR^TW^T = WW^T$$

We can immediately see that the likelihood $p(x_i|\theta) = \mathcal{N}(x_i|\mu, \Psi + WW^T)$ remains the same.

B

MPG (Fuel Efficiency)



price

Discussion:

the horizontal axis represents price, corresponding to the features labeled "dealer" and "retail", with expensive cars on the right.

The vertical axis represents MPG, fuel efficiency versus size: heavy vehicles are less efficient and are higher up, whereas light vehicles are more efficient and are lower down.

2 PCA and KPCA

A. Consider a linear-Gaussian latent-variable model having a latent space distribution $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{x}|\mathbf{0}, \mathbf{I})$ and a conditional distribution for the observed variable $p(\boldsymbol{x}|\boldsymbol{z}) = \mathcal{N}(\boldsymbol{x}|\mathbf{W}\boldsymbol{z} + \boldsymbol{\mu}, \boldsymbol{\Phi})$ where $\boldsymbol{\Phi}$ is an arbitrary symmetric, positive definite noise covariance matrix. Now suppose that we make a non-singular linear transformation of the data variables $\boldsymbol{x} \rightarrow \mathbf{A}\boldsymbol{x}$, where $\mathbf{A}$ is a $D \times D$ matrix. If $\boldsymbol{\mu}_{ML}$, $\mathbf{W}_{ML}$ and $\boldsymbol{\Phi}_{ML}$ represent the maximum likelihood solution corresponding to the original untransformed data, show that $\mathbf{A}\boldsymbol{\mu}_{ML}$, $\mathbf{A}\mathbf{W}_{ML}$ and $\mathbf{A}\boldsymbol{\Phi}_{ML}\mathbf{A}^T$ will represent the corresponding maximum likelihood solution for the transformed data set. Finally, show that the form of the model is preserved in two cases:

(a) $\mathbf{A}$ is a diagonal matrix and $\boldsymbol{\Phi}$ is a diagonal matrix. This corresponds to the case of factor analysis. The transformed $\boldsymbol{\Phi}$ remains diagonal, and hence factor analysis is covariant under component-wise re-scaling of the data variables

(b) $\mathbf{A}$ is orthogonal and $\boldsymbol{\Phi}$ is proportional to the unit matrix so that $\boldsymbol{\Phi} = \sigma^2\mathbf{I}$. This corresponds to probabilistic PCA. The transformed $\boldsymbol{\Phi}$ matrix remains proportional to the unit matrix, and hence probabilistic PCA is covariant under a rotation of the axes of data space, as is the case for conventional PCA

B. For the data set given at this link, Compute the first 8 kernel principal component basis functions. Use RBF kernel with $\sigma^2 = 0.1$

A. (a)

A is a diagonal matrix

$$A_{D\times D} = \begin{bmatrix} A_{11} & 0 & \cdots & 0 \\ 0 & A_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A_{DD} \end{bmatrix}$$

$$\Phi_{D\times D} = \begin{bmatrix} \Phi_{11} & 0 & \cdots & 0 \\ 0 & \Phi_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \Phi_{DD} \end{bmatrix}$$

The transformed $\Phi \rightarrow A\Phi A^T$

$$A\Phi A^T = \begin{bmatrix} A_{11} & 0 & \cdots & 0 \\ 0 & A_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A_{DxD} \end{bmatrix} \begin{bmatrix} \Phi_{11} & 0 & \cdots & 0 \\ 0 & \Phi_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \Phi_{DD} \end{bmatrix} \begin{bmatrix} A_{11} & 0 & \cdots & 0 \\ 0 & A_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A_{DD} \end{bmatrix}$$

$$= \begin{bmatrix} A_{11}\Phi_{11}A_{11} & 0 & \cdots & 0 \\ 0 & A_{22}\Phi_{22}A_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A_{DD}\Phi_{DD}A_{DD} \end{bmatrix}$$

$A\Phi A^T$ is diagonal

Therefore, FA is covariant under component-wise rescaling of the data variables.

b.

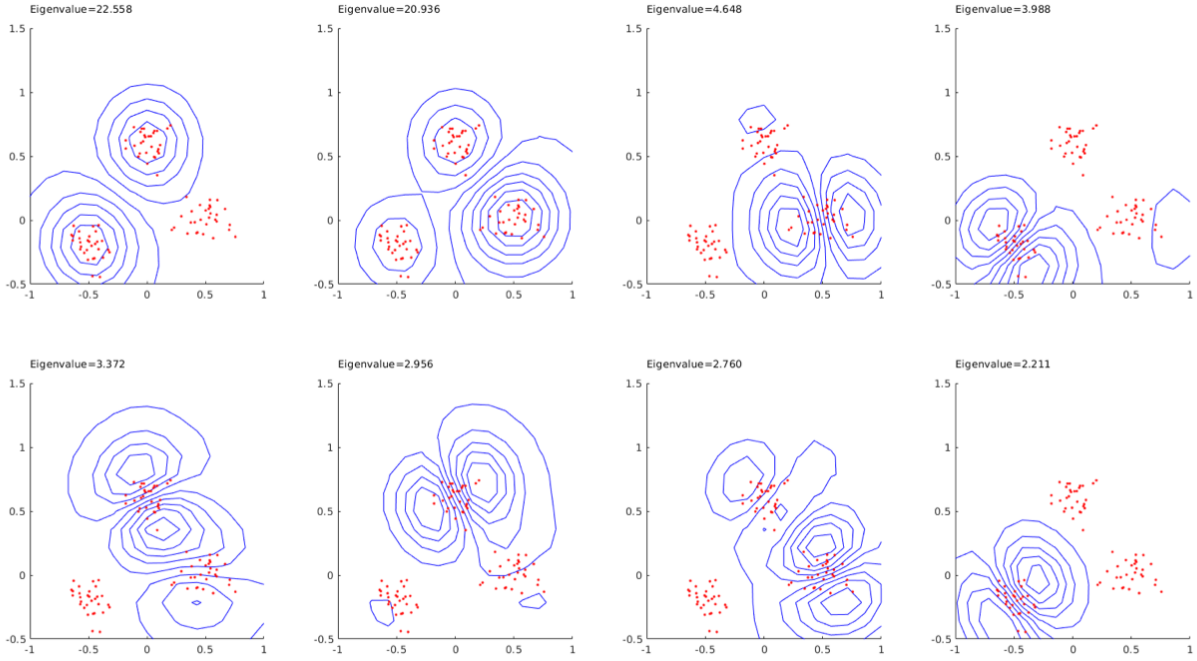
A is orthogonal $\quad AA^T = I$

$$\Sigma = \sigma^2 I$$

$$A\Sigma A^T = A\sigma^2 I A^T = \sigma^2 A I A^T$$
$$= \sigma^2 A A^T$$
$$= \sigma^2 I$$

$A\Sigma A^T$ is proportional to the unit matrix.

B.


Eigenvalue=22.558
Eigenvalue=20.936
Eigenvalue=4.648
Eigenvalue=3.988
Eigenvalue=3.372
Eigenvalue=2.956
Eigenvalue=2.760
Eigenvalue=2.211

3 Independent Component Analysis

12.6

A. Suppose that two variables z_1 and z_2 are independent. (1) Show that the covariance matrix between these variables is diagonal. Now consider two variables y_1 and y_2 in which $-1 \leq y_1 \leq 1$ and $y_2 = y_1^2$. Write down the conditional distribution $p(y_2|y_1)$ and show that this is dependent on Y_1. Now show that the covariance matrix between these two variables is again diagonal. This counter-example shows that zero correlation is not a sufficient condition for independence.

B. For some noisy observations of a 4D signal, use ICA to reconstruct the signal. The data set can be found at this link.

Figure 12.20

A. (1) z_1 and z_2 are independent.

$$P(z_1, z_2) = P(z_1)\,P(z_2)$$

$$E(z_1 z_2) = E(z_1)\,E(z_2)$$

$$\text{corr}(z_1, z_2) = \frac{E[(z_1 - E[z_1])(z_2 - E[z_2])}{\sqrt{\text{var}[z_1]}\sqrt{\text{var}[z_2]}} = 0$$

the covariance matrix between these variables is diagonal.

(2) $-1 \leq y_1 \leq 1$, assume uniform distribution. $P(y_1) = \frac{1}{2}$

$y_2 = y_1^2$, $P(y_2, y_1) = \frac{1}{y_1^2}$, $E[y_1] = 0$

$$P(y_2|y_1) = \frac{P(y_2, y_1)}{P(y_1)} = \frac{2}{y_1^2}, \text{ dependent on } y_1$$

$$\text{corr}(y_1, y_2) = \frac{E[(y_1 - E[y_1])(y_2 - E[y_2])]}{\sqrt{\text{var}[y_1]}\sqrt{\text{var}[y_2]}}$$

$$\text{corr}(y_1, y_2) = \frac{E[y_1 (y_2 - E(y_2))]}{\sqrt{\text{var}[y_1]}\sqrt{\text{var}[y_2]}}$$

$$= \frac{E[y_1 y_2 - y_1 E[y_2]]}{\sqrt{\text{var}[y_1]}\sqrt{\text{var}[y_2]}}$$

y_2 is dependent on y_1

$$y_1 y_2 = y_1 y_1^2 = y_1^3$$

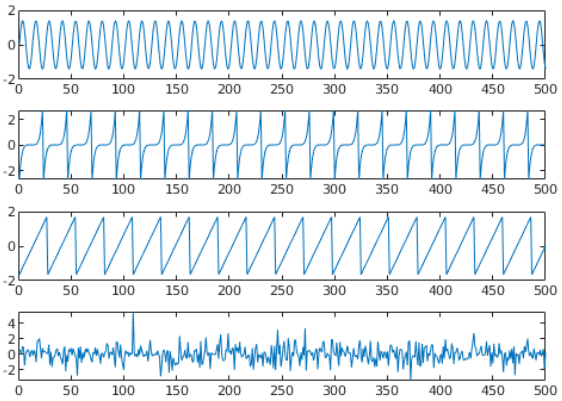
$$y_1 E[y_2] = y_1 y_1^3 = y_1^3$$

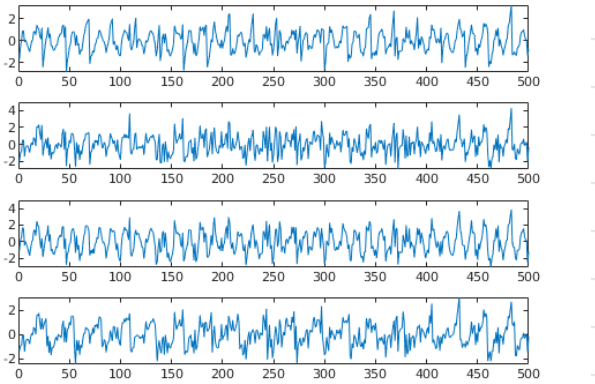
$$E[y_1 y_2 - y_1 E[y_2]] = 0$$

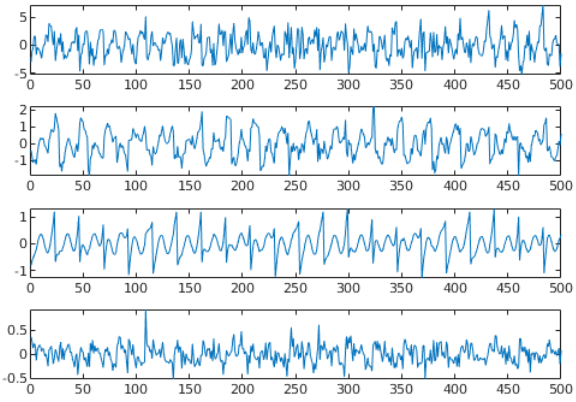
$$\text{corr}(y_1, y_2) = 0$$

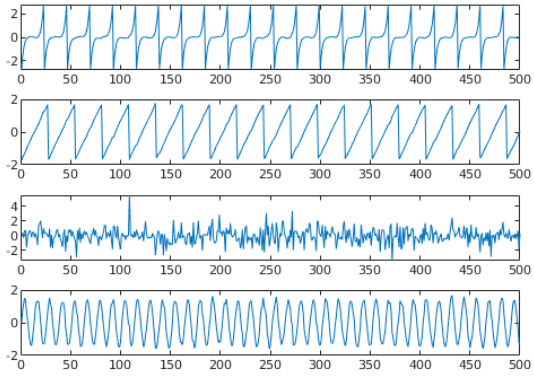
the corvariance matrix between y_1, y_2 is again diagnol.

B.









4 LASSO

13.5

A. Recall that the LASSO objective is to minimize $RSS(\beta) + \lambda \sum_j |\beta_j|$ whereas the ridge regression objective is to minimize $RSS(\beta) + \lambda \sum_j ||\beta_j|||_2^2$.

(a) **True or False:** LASSO solutions result in sparsity in the regression coefficients. Explain.

(b) **True or False:** Ridge Regression solutions result in sparsity in the regression coefficients. Explain

(c) **True of False:** As we increase λ in LASSO, we expect the number of variables in our solution to increase. Explain.

(d) **True or False:** It is possible to achieve the least-squares solution using a LASSO objective. Explain

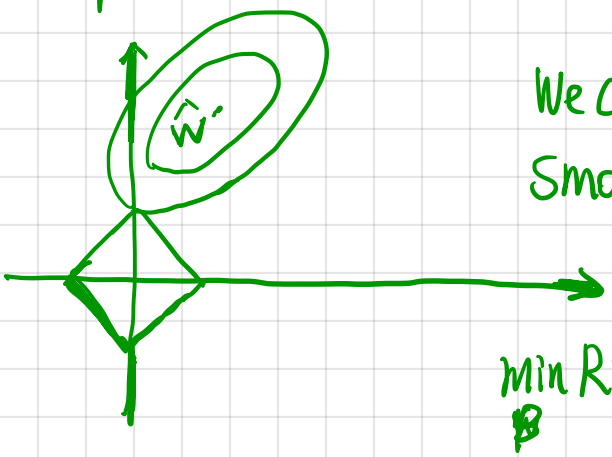
(e) **True or False:** Given any two LASSO solutions corresponding to λ_1 amd λ_2 with $\lambda_2 > \lambda_1$ and the same support for these two solutions, it is possible to write out a closed-form expression for all solutions corresponding to λ with $\lambda_1 < \lambda < \lambda_2$. Explain.

B. Generate sparse signal $\boldsymbol{w}^*$ of size $D = 4096$, consisting of 160 randomly placed ± 1 spikes. Next generate a random design matrix $\mathbf{X}$ of size $N \times D$ where $N = 1024$. Finally, generate a noisy observation $\boldsymbol{y} = \mathbf{X}\boldsymbol{w}^* + \boldsymbol{\epsilon}$ where $\epsilon_i = \mathcal{N}(0, 0.01^2)$. Estimate $\boldsymbol{w}$ from $\boldsymbol{y}$ and $\mathbf{X}$. Estimate $\boldsymbol{w}$ by using LASSO using $\lambda = 0.1\lambda_{max}$. Plot the results and discuss.

p58, slides

A: (a) True

Explain:



LASSO: $RSS(\beta) + \lambda \Sigma_j |\beta_j|$

We can re-write this as a constrained but smooth objective (a quadratic function with linear constrains)

$\min_{\beta} RSS(\beta)$ s.t. $\Sigma_j |\beta_j| \leq B$

Contour of the RSS objective function of LASSO

Where B is the upper bound on the l_1-form of weight

From the theory of constrained optimization, we know that the optimal solution occurs at the point where the lowest level set of the objective function intersects the constraint surface.

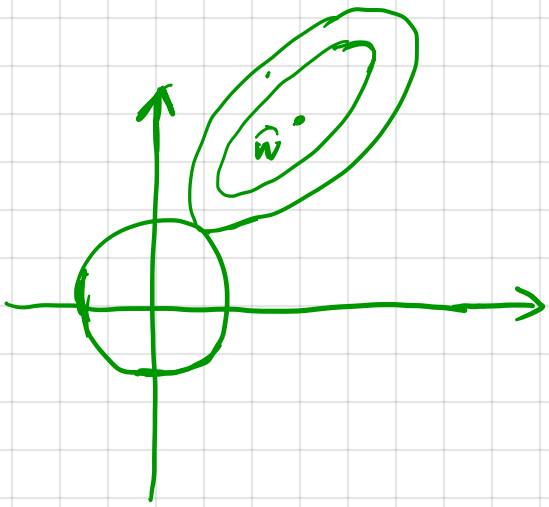
It should be geometrically clear that as we relax the constrain B, we "grow" the ℓ_1 ball until it meets the objective.

The corners of the ball are more likely to intersect the ellipse than one of the sides, especially in high dimensions, because the corners stick out more.

The corners correspond to sparse solutions, which lie on the coordinate axes.

Therefore, LASSO solutions result in sparsity in the regression coefficients.

(b) False



$\min_{\beta} RSS(\beta)$. s.t. $\sum_j \|\beta_j\|_2^2 \leq B$

When we grow the l_2 ball, it can intersect the objective at any point, there are no corners, so there is no preference for sparsity.

(c) False

As we increase λ, the solution vector $\hat{w}(\lambda)$ will tend to get sparser.

(d) True

Set $\lambda = 0$

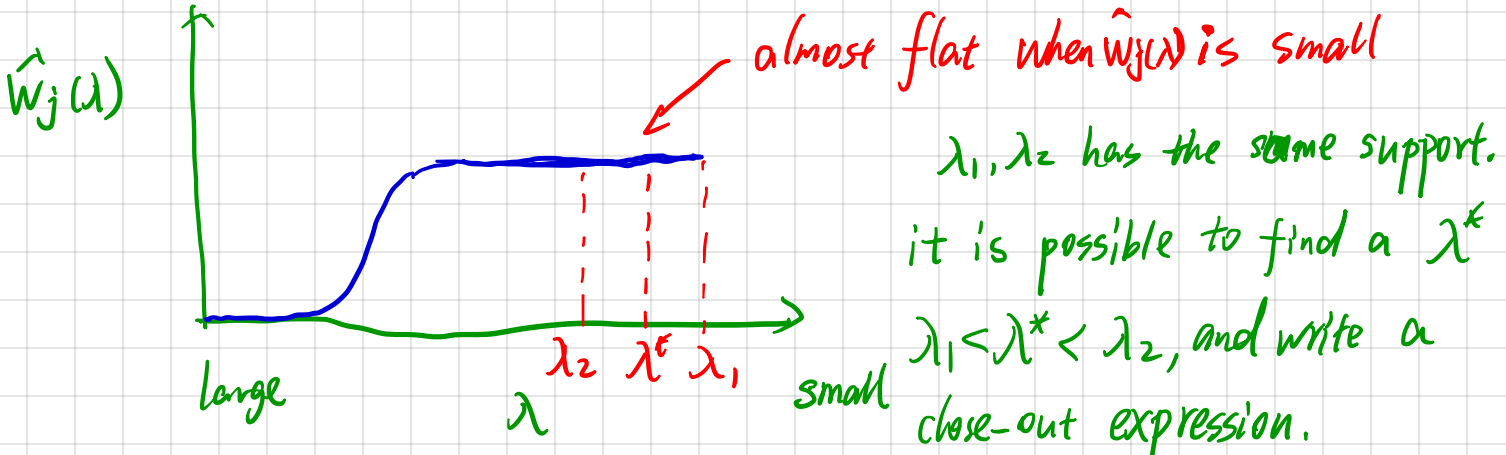
LASSO = RSS(β)

(e) True.

Explain:

from the result of regularization path.

$\hat{w}_j(\lambda)$ vs λ

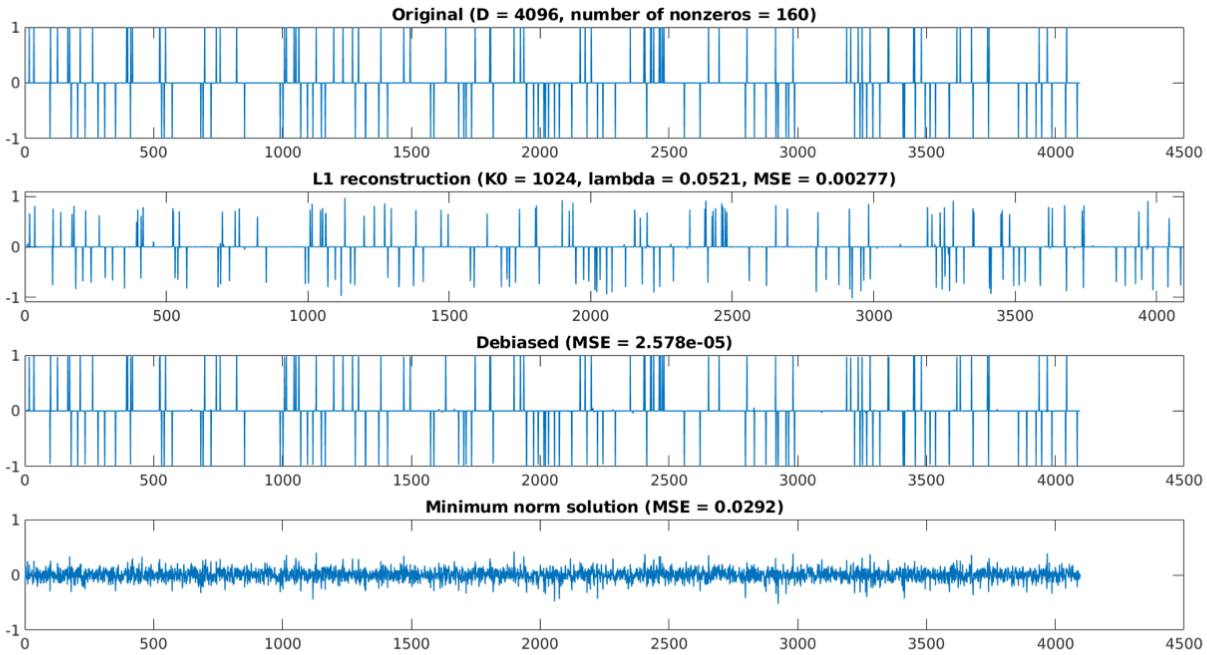


λ_1, λ_2 has the same support.

it is possible to find a λ^*

$\lambda_1 < \lambda^* < \lambda_2$, and write a close-out expression.

B.


Original (D = 4096, number of nonzeros = 160)
L1 reconstruction (K0 = 1024, lambda = 0.0521, MSE = 0.00277)
Debiased (MSE = 2.578e-05)
Minimum norm solution (MSE = 0.0292)
1
0
-1
0
500
1000
1500
2000
2500
3000
3500
4000
4500

5 Automatic relevance determination and compressed sensing

13.7

A. Automatic relevance determination is another method that results in sparse solution.

(a) Explain (with equations and diagram) how ARD works and why it results in sparse solution.

(b) How is ARD connected to the MAP estimate?

B. Suppose we have an image which is corrupted in some way, e.g., by having text or scratches sparsely superimposed on top of it. We might want to estimate the underlying clean image. This is called image inpainting. One can use similar techniques for image denoising. One way to address this problem is to use compressed sensing

Write a code for image denoising using compressed sensing. Use it for the image provided at this link.

A. (a)

$\alpha_j = \frac{1}{\sigma_j^2}$ weight precision

$\beta = \frac{1}{\sigma^2}$ measurement precision

$$p(y|x, w, \beta) = \mathcal{N}(y|w^T x, 1/\beta)$$

$$p(w) = \mathcal{N}(w|0, A^{-1}), \quad A = \text{diag}(\alpha)$$

The marginal likelihood

$$p(y|X, \alpha, \beta) = \int \mathcal{N}(y|Xw, \beta I_N)\, \mathcal{N}(w|0, A)\, dw$$

$$= \mathcal{N}(y|0, \beta I_N + X A^{-1} X^T)$$

$$= (2\pi)^{-\frac{N}{2}} |C_\alpha|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} y^T C_\alpha^{-1} y\right)$$

where $C_\alpha \triangleq \beta^{-1} I_N + X A^{-1} X^T$

$$\ell(\alpha, \beta) \triangleq -\frac{1}{2} \log p(y \mid X, \alpha, \beta)$$

$$\triangleq \log |C_\alpha| + y^T C_\alpha^{-1} y$$

$$\triangleq \log |C_\alpha| + y^T C_\alpha^{-1} y + \sum_j (a + \log \alpha_j - b\alpha_j) + d \log \beta - d\beta.$$

optimize $\hat{\alpha}, \beta$.

once we have estimated α and β, we can compute the posterior over the parameter

$$p(w \mid D, \hat{\alpha}, \hat{\beta}) = \mathcal{N}(\mu, \Sigma)$$

$$\Sigma^{-1} = \hat{\beta} X^T X + A$$

$$\mu = \hat{\beta} \Sigma X^T y$$

The fact that we computer $p(w \mid D, \hat{\alpha}, \beta)$, while simultaneously encouraging sparsity.

from Figure 13.20. (Murphy's book)

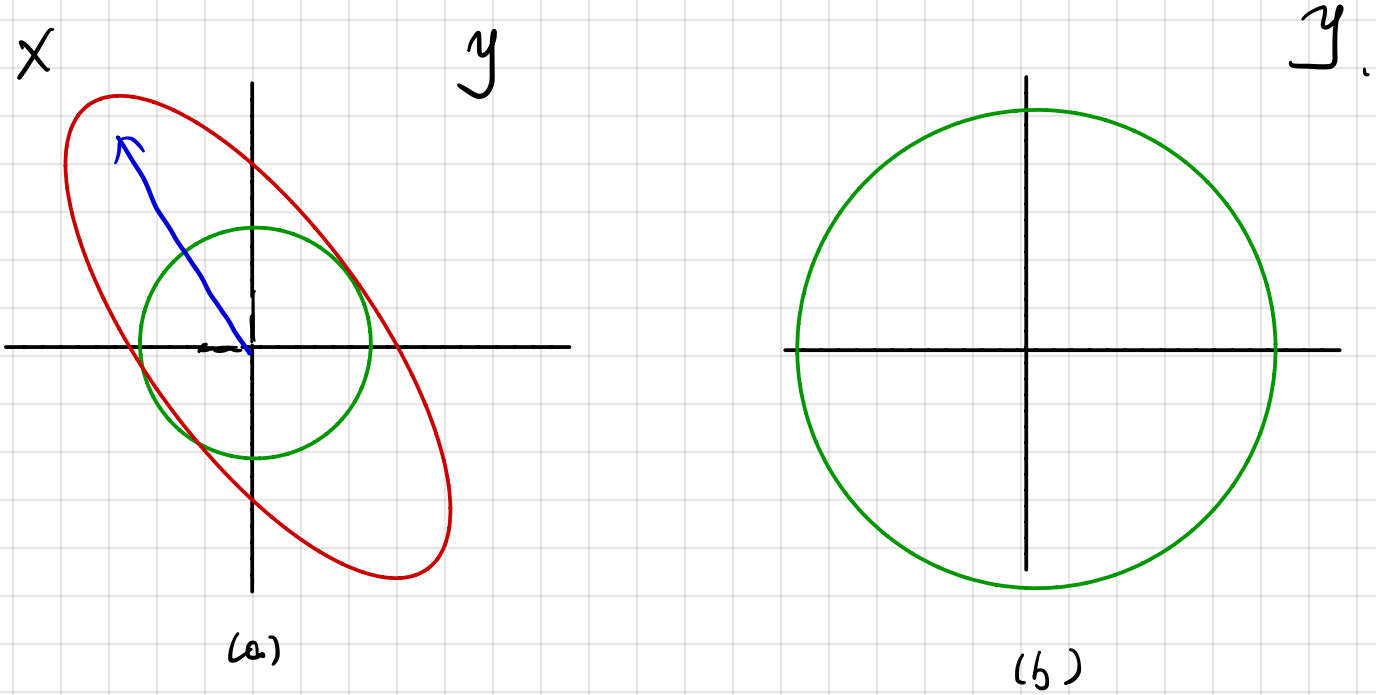


Illustration of why ARD results in sparsity. The vector of input x does not point towards the vector of outputs y, so the feature should be removed. (a) For finite α, the probability density is spread in directions away from y. (b) when $\alpha = \infty$, the probability density at y is maximized.

The marginal likelihood is

$$P(Y|X,\alpha,\beta) = \mathcal{N}(Y|0,C)$$

$$C = \frac{1}{\beta} I + \frac{1}{\alpha} XX^T$$

If α is finite, the posterior will be elongated along the direction of X.

However, if $\alpha = \infty$, $C = \frac{1}{\beta} I$, C is spherical.

If $|C|$ is held constant, the latter assigns higher probability density to the observed response vector y, so this is the preferred solution.

(b)

the effective prior $p(w|\hat{\alpha})$ is non-factorial, and further it depends on data D and α^2.

$$\hat{w}^{ARD} = \arg\min_{w} \beta \|y - Xw\|_2^2 + g_{ARD}(w)$$

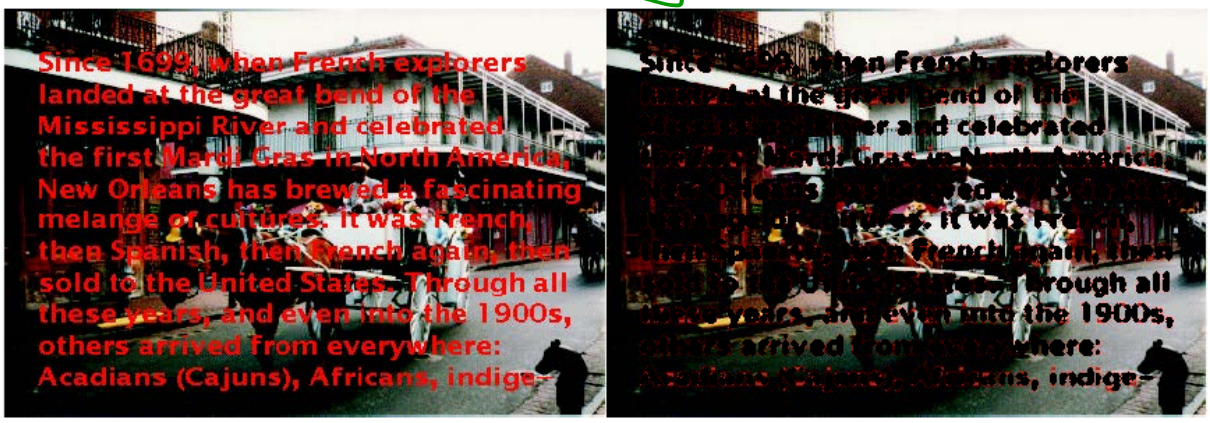
$$g_{ARD}(w) \hat{=} \min_{\alpha \geq 0} \sum_j \alpha_j w_j^2 + \log|C_\alpha|$$

ARD can be viewed as the MAP estimation problem.

MAP estimation with non-factorial priors is strictly better than MAP estimation with any possible factorial prior.

B.

original → remove red text



then use OMP to recover the image.

remove the red text

RGB. if $R \geq 180$, $G \leq 60$, $B \leq 60$,

the $R=0$, $G=0$, $B=0$

the I use omp to recover the image.

the code is from: github chongyangtao/Color-Image-Inpainting.